JCAI

Journal of Cognition and Artificial Intelligence (JCAI)
2025, Vol. 01, No. 2, PP 13-24
https://doi.org/10.33322/jcai.xxx

AI-Driven Prediction of Lung Cancer Survival for Clinical and Strategic Healthcare Decision-Making

Hira Farman[1], Syed Asad Mashhadi[2], Bilal Zafar[1], Muhammad Nafees[3], and Alisha Farman[1]

[1]*Department of Computer Science, Iqra University, Karachi, Pakistan*
[2]*Department of Business Administration, Iqra University, Karachi, Pakistan*
[3]*Department of Computer Science, Karachi Institute of Economics and Technology, Karachi, Pakistan*

Correspondence Author: Hira Farman (hira.farman@iqra.edu.pk)

Received February xx, 202x; Revised March xx, 202x; Accepted May xx, 202x

Abstract

Lung cancer is a common and deadly cancer all over the world, and early and proper diagnosis is necessary to enhance the lives of the patients. The effect of traditional methods of diagnosing a patient is that they do not necessarily reflect subtle patterns in the data of patient, which may restrict their ability to provide efficient clinical decisions. The latest developments in Artificial Intelligence (AI) and Machine Learning (ML) make it possible to discover predictive relationships and conduct a broad analysis of clinical characteristics. The analysis of a publicly available Kaggle lung cancer dataset, which included patient characteristics (age, gender, smoking status, and symptoms), was conducted using ORANGE data-mining platform in the study. The chosen predictive target was survival (Survived/Not Survived) which was used to investigate the relationship between clinical features and patient outcomes. The Decision Tree had the best performance among the assessed models (CA = 0.931, AUC = 0.983, F1 = 0.928, Precision = 0.929, Recall = 0.931, MCC = 0.790). RF also reported good scores (CA = 0.912, AUC = 0.988, F1 = 0.903, Precision = 0.919, Recall = 0.912, MCC = 0.731), whereas kNN has given mediocre results. Gradient Boosting (CA = 0.781) and Naive Bayes (CA = 0.780) had much weaker scores especially on F1-score and MCC. All in all, the results show that AI-driven predictive modeling, particularly interpretable predictive models like Decision Trees and Random Forests, can help clinicians and healthcare policymakers to determine the most significant predictors to use in making a diagnosis, treatment plan, and even strategic healthcare management.

Index Terms: *Body Mass Index, Cancer Stage, Hypertension, Machine Learning, and Smoking Status.*

I. INTRODUCTION

Lung cancer is one of the most prevalent and deadly forms of cancer that pose a major health challenge at the global level in morbidity and mortality. Recent statistics by World Health Organization (WHO) indicate that approximately, in the world, one-fourth of all the cancer-causing diseases results in the fatal illness known as lung cancer that has very devastating impact on health care systems [1], and [2]. Late-stage diagnosis and the lack of access to the latest clinical diagnosis and treatment opportunities result in the fact that the prognosis of patients with lung cancer is frequently poor to be integrated into the effective screening programs [3]. Lifestyle, genetic, and environmental factors, including smoking and exposure to pollution, also influence the prevalence of the illness and complicate the process of forecasting [5], and [7]. Consequently, timely diagnosis and timely intervention are fundamental in improving patient outcomes and reducing the general burden of the healthcare systems. Conventionally used diagnostic techniques that are often based on invasive tests or a narrow amount of clinical data normally slow down treatment and do not give a dependable forecast of disease advancement [4]. By contrast, new trends in Artificial Intelligence (AI), specifically Machine Learning (ML) and Deep Learning (DL), provide promising solutions to eliminate these limitations [8], and [9]. CNN models VGG and ResNet have shown better results in the analysis of CT and histopathological data [10]. Likewise, most classical ML algorithms such as Support Vector Machines (SVM), Logistic Regression, Random Forest, and Naive Bayes have proven to be successfully used in prediction tasks in oncology [14]. Furthermore, the ensemble and hybrid methods have demonstrated great capability in risk prediction and trend estimation [8], and [9].

The availability of datasets to the general population such as the Kaggle Lung Cancer Dataset is an effective source of developing and testing predictive models based on demographic, lifestyle and other aspects of clinical characteristics [12]. Besides clinical uses, AI based analytics also enhance healthcare management by enhancing patient stratification, resource allocation and cost effectiveness [5], and [14]. This study employs the machine learning techniques on the lung cancer data on the Kaggle dataset to evaluate the predictive accuracy as well as help in clinical decision-making and strategic healthcare management.

A. Study Objective

The goal of this endeavor is to use a publicly accessible Kaggle dataset to create and assess AI/ML predictive models for lung cancer survival.

The work aims to:

a) Support clinical practice by predicting patients' survival status (Survived/Not Survived) early and accurately based on clinical, lifestyle, and demographic characteristics; and
b) Help healthcare administrators make strategic decisions by connecting survival predictions to planning, resource allocation, and cost-effective management of oncology services.

B. Contribution of this Study

The main contributions of this study are as follows:

- Integration of survival prediction + strategic healthcare decision-making.
- Using feature ranking (Chi-square + ReliefF).
- Using interpretable ML models for clinical transparency.
- Using a large-scale public dataset for reproducible benchmarking.
- Implications for hospital planning, insurance, and triage.

II. BACKGROUND STUDY

According to authors several machine learning algorithms such as Naive Bayes, Support Vector Machines (SVM), Logistic Regression, and Artificial Neural Networks (ANN) have been extensively applied in detecting and diagnosing lung cancer [1]. Their study formed the foundation of the integration of AI into healthcare decision making by explaining the comparative benefits and costs of various algorithms.

Researchers examined the frameworks of Machine Learning (ML) to classify the pulmonary nodules and the authors were concerned with how they will improve the accuracy of the diagnosis and reduce unnecessary clinical follow-ups [2]. They also noted issues with dependability and validation or impediments to clinical application, based on the UCI lung cancer data. Research analysts used the following classifier: SVM, K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNN) [3]. They found out that SVM was often more accurate than CNN and KNN, and can be applied in the early detection task. It was analyzed that such methods as SVM, Naive Bayes, Decision Trees, and Logistic Regression could be used to forecast lung cancer [4]. Their results explained the importance of the evaluation of algorithms in enhancing early diagnosis and survival. A study indicated that lung cancer is not detected at its initial stage, which also leads to the high mortality rate of this condition [5]. In order to facilitate early detection, classification, and evaluation of the malignancy, they emphasized the increasing importance of the Computer Aided Diagnostic (CAD) systems that are based on the machine learning, deep learning, and image processing techniques applied to the UCI data. To allow distinguishing between the carcinogenic and the non-carcinogenic cases, a study utilized classifiers such as Radial Basis Function (RBF) network in WEKA on the UCI data [6]. This study has not only proven the applicability of machine learning in classification, but it has also proven smoking to be a significant risk factor. In order to find the genetic leads concerning lung cancer, a research study employed the eco- genomics, the data on expression of the genes of the repositories like the Kent Ridge Biomedical Repository [7]. It was discovered that genetic and environmental elements perform better to enhance performance in prediction.

Random Forest was among other ensemble based methods of identifying the high risk persons that was identifies and tested [8]. These techniques according to their research have the potential of creating more accurate and reliable lung cancer risk prediction models in order to interfere early. Analysts have used techniques such as Support Vector Regression, Backpropagation neural network and Long Short Memory (LSTM) to predict the presence of lung cancer on a large scale basis [9]. Their study showed how machine learning can be used in trend analysis and predictive epidemiology. Examiners developed deep learning models based on relevant CNN architectures such as AlexNet, LeNet, and VGG-16 that were applied on CT images [10]. Their results indicated that early diagnosis was possible since their CNN-based models showed the feasibility of distinguishing between normal and aberrant lung tissues. A research based study used image preprocessing, segmentation, and feature extraction with classifiers which include SVM, Random Forest and ANN [11]. The hybrid pipeline was effective in separating benign and prevalent tumors, which proves the usefulness of hybrid pipelines in machine learning and image processing.

Researchers expanded the use of machine learning to forecast a wide range of cancer such as prostate, breast, and lung cancers [12]. They applied algorithms such as SVM, KNN, CART and random forest to classify the tumors into malignant, benign and high risk groups, indicating the diversity of machine learning in oncology.

A study utilized text based symptom datasets using classic machine-learning classifiers to predict lung-cancer [13]. Nevertheless, the work was later withdrawn because of methodological inconsistencies, and even when it was published, the study failed to cover the survival analysis, and even in its initial form, it failed to cover the structured clinical characteristics, including comorbidities and forms of treatments. Examiners predicted the survival of lung-cancer patients that were trained on the SEER clinical registry using supervised machine-learning models, such as Gradient Boosting, Decision Trees, and ensemble classifiers [14]. Even though their study showed good predictive work, it failed to project model outputs to the healthcare management, operational planning, or resource allocation. Researchers were also interested in the CT-image-based diagnosis with the help of deep-learning architectures CNN, VGG, and ResNet [15]. Although they worked well in classifying an image, their method did not model the outcome of survival and did not also include demographic, lifestyle, or treatment-related factors. Research Scientists studied supervised ML methods in non-medical setting (telecom churn forecasting) where the authors showed their prior experience with the interpretable models, but the study was not related to clinical survival modeling or medical decision-making [16]. Using structured meteorological datasets, In another study which assessed a number of machine-learning models for precipitation forecasting in

Australia, concentrating on comparative performance across various geographic regions [17]. The study showed that rainfall variability can be accurately captured using ML techniques; In order to facilitate smart-agriculture planning in Pakistan, some researchers suggested a data-driven machine learning strategy for rainfall prediction [18]. The study examined soil-moisture and meteorological variables using supervised learning techniques.

In order to forecast survival outcomes for patients with Non-Small Cell Lung Cancer (NSCLC) who experienced brain metastases, Researchers used machine-learning approaches. Their study demonstrated the promise of AI for outcome prediction in late-stage oncology by stratifying patients by prognosis using supervised machine learning models [19]. Nevertheless, the study was restricted to a specific clinical subgroup (NSCLC with brain metastases) and relied on specific clinical factors, which limits its applicability to larger populations of lung cancer patients. Furthermore, neither operational planning nor healthcare management viewpoints were included in the study.

To predict lung cancer survival, researchers suggested a hybrid approach that combines statistical survival models and machine learning classifiers [20]. Their research demonstrates the value of combining Machine Learning (ML) with conventional statistical techniques like Cox-based survival modeling. Nevertheless, interpretability, feature-ranking, and integration with clinical decision-making workflows were not investigated, and the investigation was carried out on a small dataset. Additionally, the analysis did not apply its predictions to administrative or strategic aspects of healthcare systems.

A thorough assessment of machine learning and deep learning models for lung cancer level/stage classification was carried out by researchers [21]. Models with competitive accuracy on structured feature sets included Random Forest, SVM, CNNs, and hybrid techniques. However, the study did not look at how model outputs could help with resource allocation, patient prioritizing, or medium-term healthcare planning; instead, it concentrated on stage categorization rather than survival prediction. Furthermore, there was no examination of operational deployment or ethical issues in the paper.

The current research, in contrast, fills several of the gaps present in these papers; **(i)** by operating on structured clinical and lifestyle characteristics, rather than only imaging or text data, **(ii)** by specifically targeting survival prediction over detection alone, **(iii)** by relying on interpretable ML models, which could be used in clinical transparency, and **(iv)** by connecting survival predictions to medium-term hospital planning and strategic healthcare decision-making, an area that has not been adequately explored in prior studies of lung-cancer prediction. The literature review reveals that the machine learning and deep learning models were already popular in predicting, detecting, and analyzing lung cancer survival. Although such traditional classifiers like SVM, Naive Bayes and Logistic Regression have consistently performed on both clinical and symptom based data, the most recent development on CNNs, ensemble and hybrid pipelines has shown to be successful at high rates on image classification, risk and trend prediction. Despite such a massive improvement, difficulties such as data reliability, validation, and clinical integration still exist. The combination of these works forms a solid base of AI solutions in the field of lung cancer diagnostics and prognosis contributing to the fact that more powerful and clinically valid predictive algorithms are built.

The Table I shows the Comparative Overview of Recent Lung Cancer AI Studies.

Table I: Comparative Overview of Recent Lung Cancer AI Studies

S. No.	Study	Dataset	Methods	Objective	Limitation / Gap
1.	Raoof et al. [1]	UCI	NB, SVM, ANN	Lung cancer prediction using classical ML	No survival prediction; no healthcare management integration
2.	Abdullah et al. [3]	UCI	SVM, KNN, CNN	Classification based on correlation-selected features	Small dataset; focus only on early detection
3.	Lynch et al. [14]	SEER	Gradient Boosting, Decision Trees, Ensembles	Survival prediction using clinical attributes	No link to strategic healthcare operations or resource planning
4.	Mamatha et al. [15]	CT Image Dataset	CNN, VGG, ResNet	Imaging-based tumor diagnosis	No clinical-feature–based survival modeling; image-only approach
5.	This Study	Kaggle	DT, RF, KNN, NB, GB	Survival prediction + strategic healthcare decision-making	Adds feature ranking + managerial implications; still needs clinical validation

III. METHODOLOGY

This paper adheres to the workflow depicted in Figure 1, with the primary stages being the data preparation, data analysis, and evaluation of the model.

A. Data Source

The data used within the research was the one at the Kaggle repository, which contains information about the patients with lung cancer publicly. It includes the attributes such as age, gender, smoking behavior and symptoms, which the attributes can be used as predictive attributes. The reason behind this was that the selection of the attributes was done based on their clinical significance in understanding the course of illness. It has 17 columns and 199,999 rows of patient characteristics, diagnoses and information about their treatment. It has three numerical features and ten categorical features (three of them are numerical and the rest are categorical features), two date features and two metadata features.

The target measure of the study is survival (Survived /Not Survived). It is important to note that there are no values that are not present in the data. The most meaningful features were mentioned in the stage of preprocessing so that they could be incorporated in the modeling step. The data was further divided into two (training and testing) to quantify the effectiveness of models applied in the research; the models applied are Naive Bayes, Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, and Decision Tree. Performance was evaluated using evaluation metrics such as accuracy, precision and recall. It was repeated in preprocessing, feature selection and model training to enhance the predictive power and reliability.

The Kaggle lung cancer dataset's size, accessibility, and organized representation of important clinical and demographic factors are what drive its utilization. In order to enable repeatability and comparison across various modeling methodologies, public benchmark datasets are frequently employed in the early phases of AI research in healthcare. Prior to clinical implementation, a number of earlier studies on cancer prediction and survival analysis also relied on public archives. We treat this study as a methodological and benchmarking contribution rather than a finalized clinical tool in accordance with this practice.

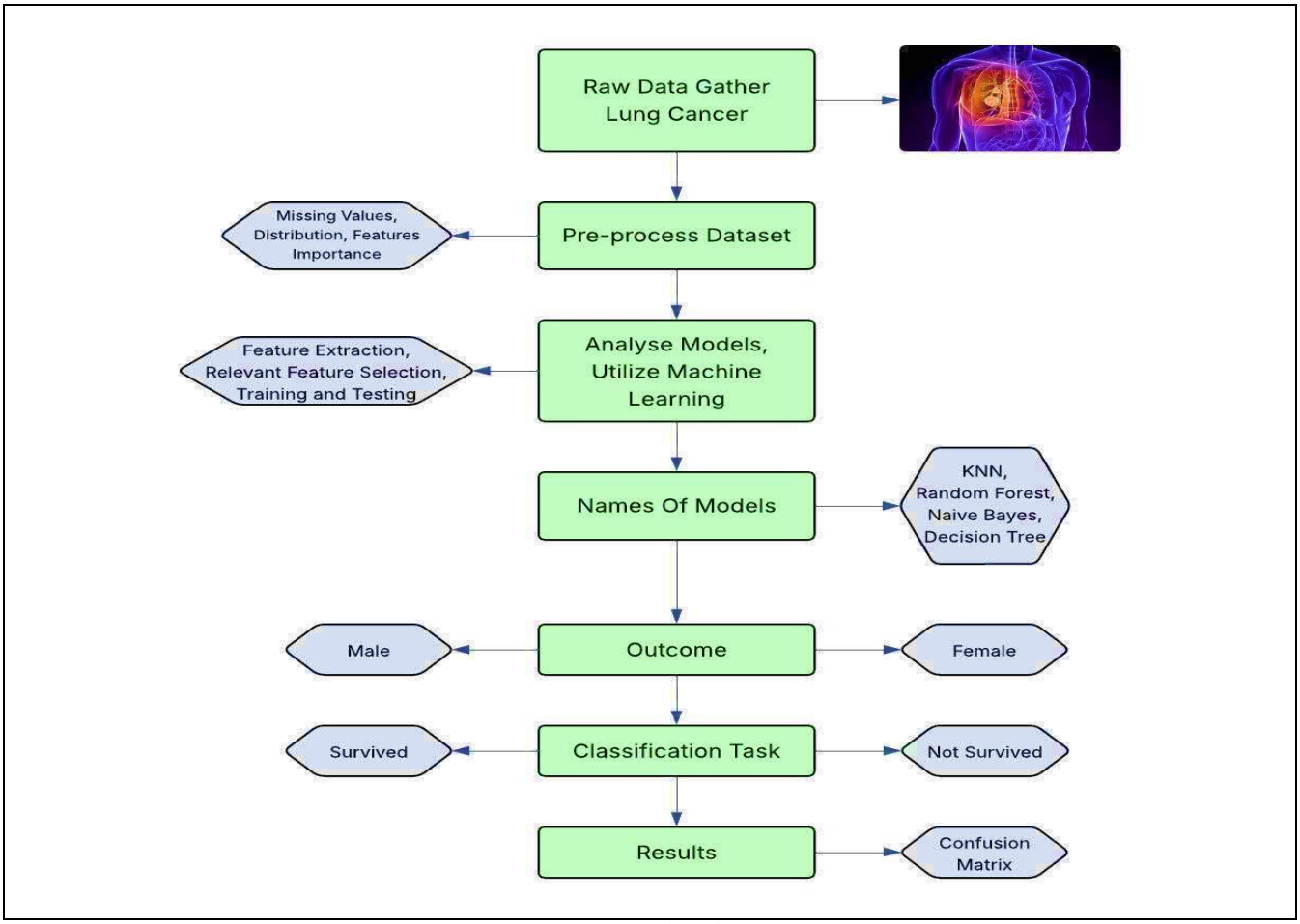

Figure 1: Workflow of the Study

B. Data Preprocessing and Preparation

The dataset was preprocessed to be of quality, consistent and suitable to be analyzed using machine learning methods. As the Kaggle data did not have any missing data, the data imputation was not necessary. The attributes analyzed in the statistical design to detect significant trends were the main features of age, smoking status, and gender. This was followed by the feature importance analysis which was conducted to give the most relevant predictors a priority hence enhancing efficiency of the model and eliminating noise. The target variable was a categorical variable which denoted survival (Survived/Not Survived). Lastly, the categorical variables like the smoking status, type of treatment and comorbidities were coded as numerical values to allow the application of machine learning algorithms. The Figure 2 shows the Data analysis using orange flow and Table II summarizes the characteristics of the Lung Cancer dataset used in this study, including its size, feature composition, target variable, Meta attributes, and the absence of missing data as defined in the Orange workspace.

We used stratified sampling to divide the dataset into training and testing sets in order to minimize bias caused by imbalance. We assessed the models using precision-recall behavior, Matthews Correlation Coefficient (MCC), accuracy, and AUC.

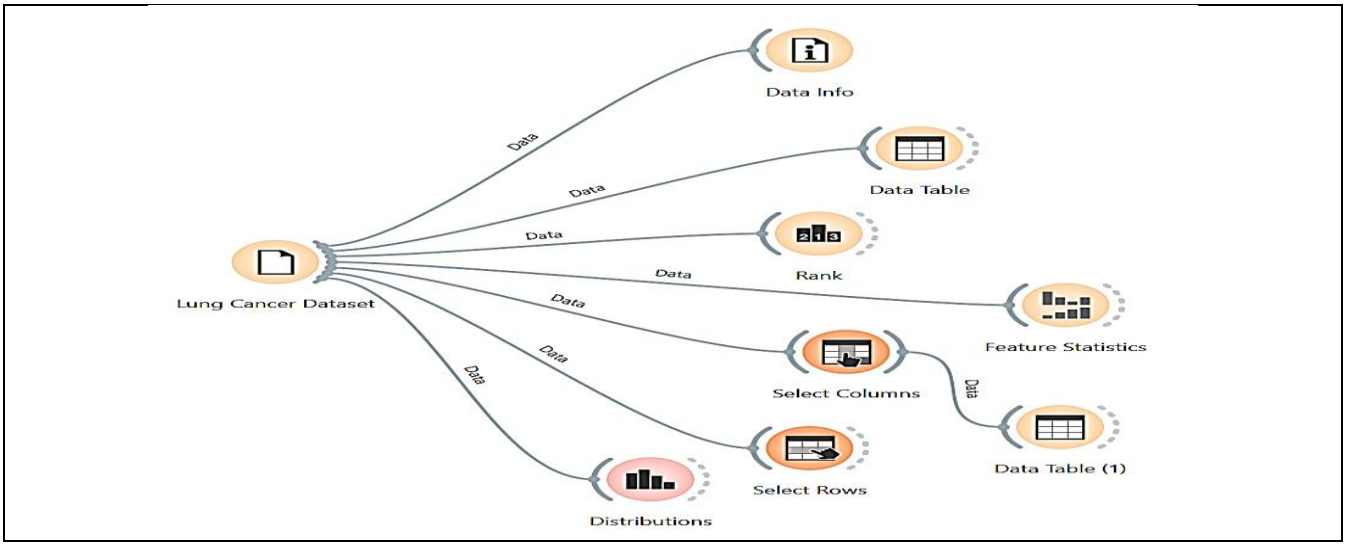

Figure 2: Dataset Analysis using Orange Flow

Table II: Orange Workspace Description

S. No.	Property	Description
1.	Dataset Name	Lung Cancer
2.	Size	199,999 Rows, 17 Columns
3.	Features	10 Categorical, 4 Numeric
4.	Target Variable	Categorical Outcome with 2 Classes
5.	Meta	2 Text Attributes
6.	Missing Data	None

C. Class Imbalance Handling

The dataset contains an uneven distribution between the Survived and Not Survived classes. Although ORANGE does not support SMOTE or advanced resampling techniques directly, the "Balance Data" widget was used to adjust class weights and reduce skew during model training. Additionally, stratified training/testing splits were applied to preserve class proportions. This ensures a more reliable evaluation, especially when combined with imbalance-robust metrics such as MCC and Precision–Recall analysis.

D. Data Features Statistics

The data set includes information on patients having lung cancer in detail covering their treatment route and their lifestyle and medical history.

Age defines the age of the patient and ID is a unique identifier. Nation of origin of every patient is specified in Country and the date of diagnosis and the date of the end of their treatment are captured in the Timeline, and the most vital lifestyle and clinical characteristics are Smoking Status, which defines the smoking history of the patient and Cancer Stage, which identifies the stage of cancer progression. There are also other health variables like the level of cholesterol in the body and the Body Mass Index (BMI). Comorbidities are in the form of binary indicators like cirrhosis, asthma, hypertension, and other cancers. Lastly, Treatment Type: This identifies the medical procedures one undergoes, i.e. chemotherapy, surgery or radiation. The Survived (Survived/Not Survived) is a categorical feature that is to be used to define the target variable in the proposed study. Taken together, these characteristics give a global perspective on patient demographics, clinical conditions, and treatment information, thus the dataset is appropriate to test machine learning models.

The overall feature statistics are shown in figures, i.e., Figure 3 to 5.

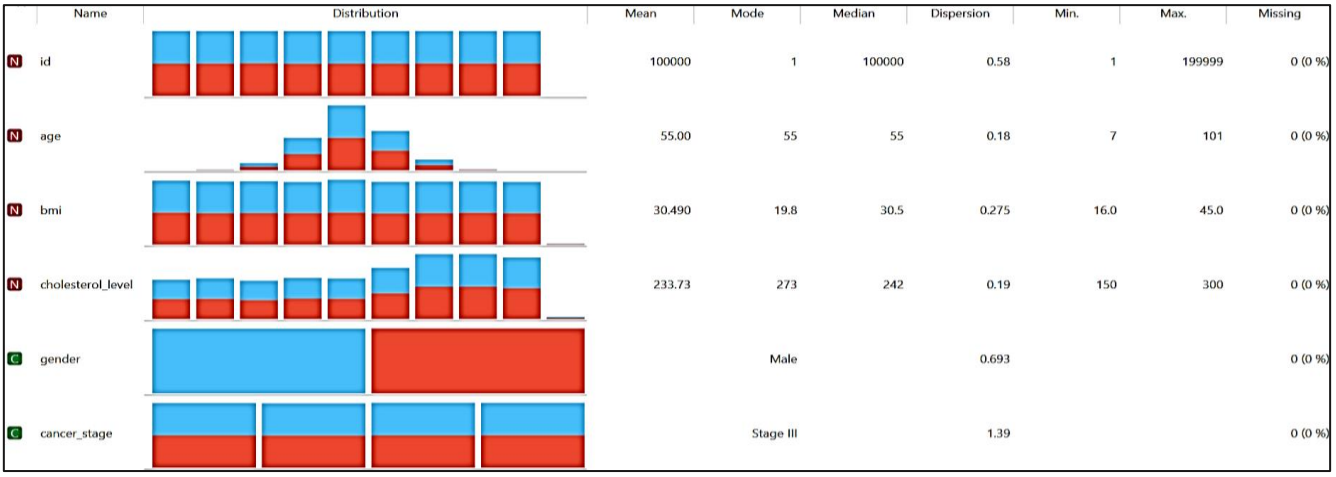

Figure 3: Feature Statistics 1

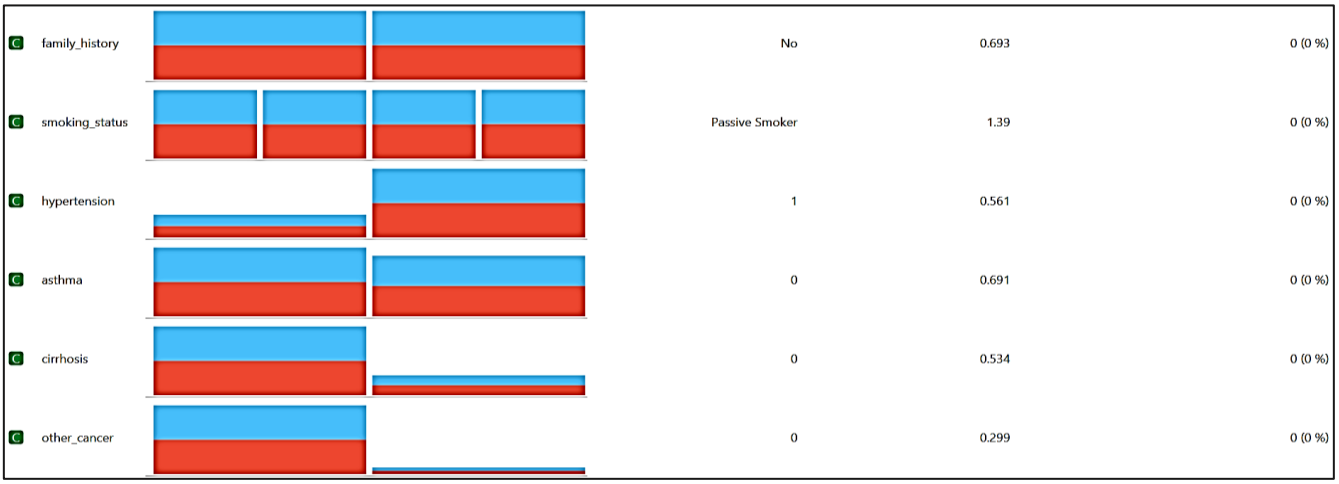

Figure 4: Feature Statistics 2

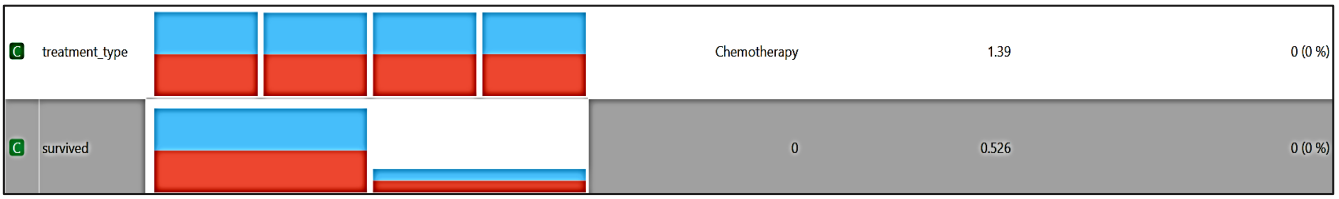

Figure 5: Feature Statistics 3

In Figure 6, a ranking analysis of the lung cancer data has been provided with two scoring processes, namely ReliefF and Chi-square (2). The predictive value of each of the characteristics in the development of lung cancer is established. The information encompasses country, condition of survival, body mass, age, form of therapy, asthma, cancer stage, cholesterol, cirrhosis, hypertension, additional cancer, smoking condition and family history. The more the higher the score the better are the correlations of each feature with the target variable, depending on the chi-square. Conversely, ReliefF values estimate the discriminative power of the features; that is, the higher the value the more the focus on the features that give the greatest classification to the classes. This ranking finds the most important indicators of the creation of correct and sound machine learning models.

		#	χ^2	ReliefF
1	C treatment_type	4	2.503	-0.006
2	C family_history	2	0.588	-0.004
3	C gender	2	1.478	-0.000
4	C other_cancer	2	0.002	0.004
5	C smoking_status	4	1.466	0.006
6	C hypertension	2	0.315	0.006
7	C asthma	2	0.501	0.006
8	N age		1.148	0.012
9	C cirrhosis	2	0.220	0.016
10	N bmi		0.252	0.018
11	C cancer_stage	4	0.516	0.022
12	N cholesterol_level		1.179	0.027

Figure 6: Rank Statistics of Features

Significantly, the ReliefF value and Chi-square values of such characteristics as survived and type of treatment use are positive and relatively high, which indicates their relevance to the outcomes predictions related to lung cancer. On the other hand, low ReliefF scores of traits may not be so important and even less probable to reduce the predictive power of the model. This ranking is necessary in order to ascertain the most influential

predictors to machine learning modeling in lung cancer studies. When it comes to a manual selection process the top five features are prioritized. The dataset also gives the gender distribution of the patients as shown by the bar chart. There are approximately 100,000 entries in each group, approximately an equal amount of male and female patients. The dataset is protected from gender bias because of its equal representation, which enables a fair assessment of predictive models for patients of both genders shown in Figure 7 and the overall data represented in Figure 8.

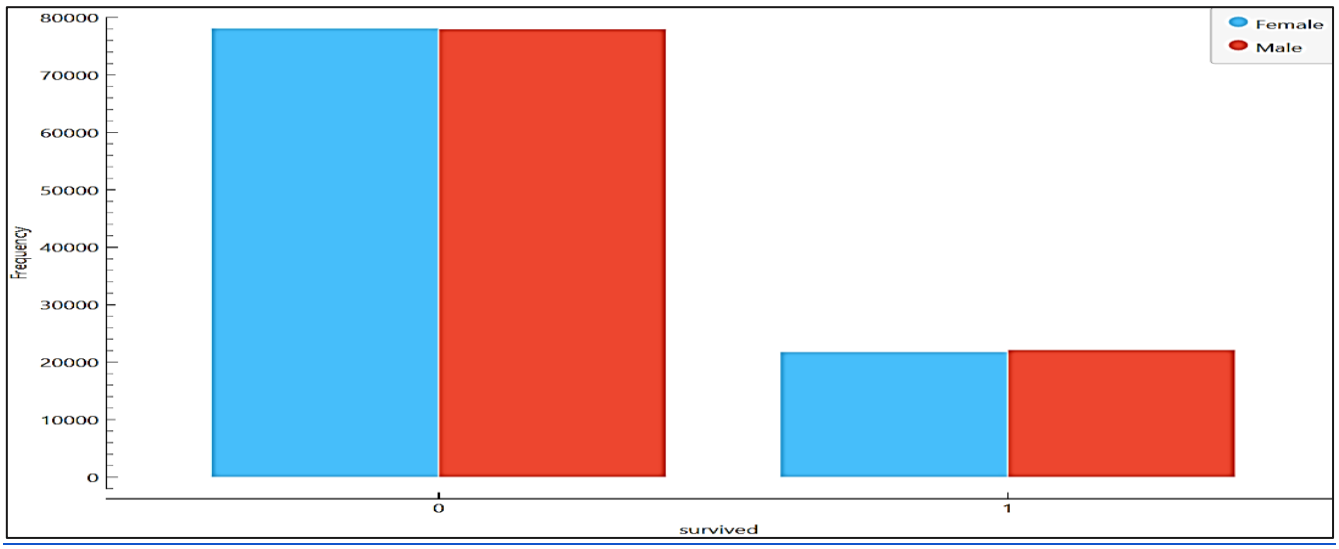

Figure 7: Distribution on Gender Target Feature

	cancer_stage	smoking_status	bmi	cholesterol_level	hypertension	asthma	cirrhosis	other_cancer	treatment_type	survived
1	Stage I	Passive Smoker	29.4	199	0	0	1	0	Chemotherapy	0
2	Stage III	Passive Smoker	41.2	280	1	1	0	0	Surgery	1
3	Stage III	Former Smoker	44.0	268	1	1	0	0	Combined	0
4	Stage I	Passive Smoker	43.0	241	1	1	0	0	Chemotherapy	0
5	Stage I	Passive Smoker	19.7	178	0	0	0	0	Combined	0
6	Stage I	Never Smoked	37.6	274	1	0	0	0	Radiation	0
7	Stage III	Passive Smoker	43.1	259	0	0	0	0	Radiation	1
8	Stage IV	Former Smoker	25.8	195	1	1	0	0	Combined	0
9	Stage III	Current Smoker	21.5	236	0	0	0	0	Chemotherapy	0
10	Stage IV	Current Smoker	17.3	183	1	0	0	1	Surgery	0
11	Stage IV	Never Smoked	30.7	262	1	1	0	0	Surgery	1
12	Stage II	Former Smoker	33.9	287	0	0	0	0	Combined	0
13	Stage II	Current Smoker	25.6	163	0	1	0	0	Chemotherapy	0
14	Stage IV	Never Smoked	26.3	174	1	1	1	0	Combined	0
15	Stage II	Former Smoker	42.7	259	1	1	0	0	Radiation	0
16	Stage IV	Passive Smoker	19.6	158	1	1	1	0	Surgery	0
17	Stage III	Former Smoker	21.7	195	1	0	0	0	Radiation	0
18	Stage II	Former Smoker	23.1	213	0	0	0	0	Combined	0
19	Stage IV	Current Smoker	43.4	251	0	1	0	1	Surgery	0
20	Stage III	Current Smoker	36.8	270	1	1	0	0	Chemotherapy	0
21	Stage I	Passive Smoker	24.6	219	1	0	1	0	Radiation	0
22	Stage III	Former Smoker	16.0	232	1	1	0	0	Radiation	1
23	Stage IV	Former Smoker	38.0	295	1	1	1	0	Surgery	0
24	Stage I	Never Smoked	38.0	287	1	0	0	1	Chemotherapy	0

Figure 8: Data Table

The study's dataset includes survival outcomes as well as clinical, lifestyle, and treatment-related data about individuals with lung cancer. Its characteristics include body mass index (BMI), cholesterol, smoking status (never smoked, passive smoker, former smoker, or current smoker), and cancer stage (Stage I–IV). There is also evidence of concomitant conditions such as cirrhosis, asthma, hypertension, and various cancers. The details of the treatment, such as chemotherapy, radiation, surgery, or a combination of these, are also covered. The goal variable is a binary representation of the patient's survival status. This dataset provides a thorough foundation for lung cancer survival prediction modeling and aids in healthcare decision-making by combining lifestyle and medical aspects.

E. Model Development

The number of machine learning methods for predicting patient survival was compared using the ORANGE data mining platform. It is worth mentioning that the use of somewhat classical machine learning models in this

experiment was not fortuitous. The predictive power of a model in its raw form is as important as its transparency and interpretability in a number of clinical applications. Random Forests and Decision Trees can be read by humans and have importance profiles, which can be discussed directly with clinicians and administrators and other machine learning models too utilized in study. Although more complex deep learning models might be more accurate in theory, they would be a black-box, and thus not be accepted in healthcare decision-making due to the traceability. Therefore, what we focus in this work is the set of interpretable models which may be adopted in practice in a real way.

The models listed below are assessed:

a) Decision Tree (DT):

This model makes step by step root-to-leaf decisions using a sequence of yes and no questions on patient characteristics that are like a tree to identify survival.

b) Random Forest (RF):

Random Forest develops a number of Decision Trees depending on the subsets of the data, and averages the results through majority voting. It is an ensemble approach that reduces overfitting, increases the generalization and produces a more accurate and more reliable forecast of patient survival.

c) Naïve Bayes (NB):

Naive Bayes is a predictive method of survival based on probability theory and under the condition that the features of patients are conditionally independent. It is straightforward but effective at medical classification since it can determine the probability of survival by adding the respective probabilities of every feature.

d) K Nearest Neighbors (KNN):

The K-Nearest Neighbors (KNN) model predicts a patient's survival based on the outcomes of the closest, most similar patients in the data. When making a choice, it considers "neighbors" who share comparable characteristics.

e) Gradient Boosting(GB):

Gradient Boosting is an ensemble based learning algorithm where the prediction models are constructed in a chain fashion with each new tree making corrections to the older ones. It integrates several weak learners, usually shallow Decision Trees, into a strong predictive model through loss optimization via gradient descent. Gradient Boosting is a good approach to predicting survival in patients because it can effectively represent the complex non-linear relationships between patient characteristics and patient survival outcomes with high precision and resistance to overfitting as long as tuned appropriately.

F. Hyperparameter Tuning and Model Settings

All the models have been set and optimized through the parameter-setting and optimization capabilities of the ORANGE data mining platform. In the case of the Decision Tree classifier, we altered the maximum depth of the tree and the minimum sample size in each tree leaf to prevent overfitting and still achieve interpretability. Random Forest model has been optimized by changing the amount of trees per random forest (between 50 and 200), the maximum depth, and the bootstrap sampling. In the case of K-Nearest Neighbors (KNN) we tried various values of. K (3-15) and distance measures. Naive Bayes made use of its default smoothing parameters as is typical in medical classification tasks. The learning rate, the number of boosting stages and the tree depth were adjusted to tune Gradient Boosting. In general, we preferred parameter settings that offered an acceptable predictive accuracy and the complexity of the model, which is required in clinical settings related to transparent and explainable models.

IV. RESULTS OF MACHINE LEARNING MODELS

These models were chosen due to their ability to handle a range of data types and their proven track record of effectiveness in classifying tasks. As shown in Figure 9, the Orange dashboard sample summarizes the performance of the evaluated models.

The Figure 9 shows the arrangement of orange widgets and how the results of the machine learning models are displayed. Such as k-Nearest Neighbors (KNN), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), and Naive Bayes (NB), were compared based on six commonly used evaluation metrics, such as Area under the Curve (AUC), Classification Accuracy (CA), F1-Score, Precision, Recall, and Matthews Correlation Coefficient (MC). The result is presented in Table III. The assessment was on precision, discriminative power, and predictive balance. The decision tree performed well on all measures, and the precision, recall, accuracy (CA), AUC, F1-score, and MCC values were 0.929, 0.931, 0.931, 0.983, 0.928, and 0.790, respectively, which is why it is the best-performing model in general. This trade-off between high precision and recall proves this to be effective in prediction as well as discrimination. The performance of the Random Forest was also very impressive, as the accuracy was 0.912, the accuracy was 0.988, the F1-score was 0.903, the precision was 0.919, the recall was 0.912, and the MCC was 0.731. Random Forest produced an outstanding discriminatory power, but its MCC was worse than the Decision Tree (0.790), which corrected the previous inconsistency. Nonetheless, RF has been a stable and resilient classifier that has always gotten good results. By contrast, weaker results were obtained with kNN, with an accuracy of 0.798, an AUC of 0.788, an F1-score of 0.757, a precision of 0.771, a recall of 0.798, and an MCC of 0.275. Naive Bayes did not perform well, with an accuracy of 0.780, an AUC of 0.507, an F1-score of 0.684, a precision of 0.609, a recall of 0.780, and an MCC of 0.000, which indicates that it has no significant predictive capability. Equally, Gradient Boosting reported low predictive validity, with an accuracy of 0.781, an AUC of 0.535, an F1-score of 0.684, a precision of 0.829, a recall of 0.781, and an MCC of 0.009.

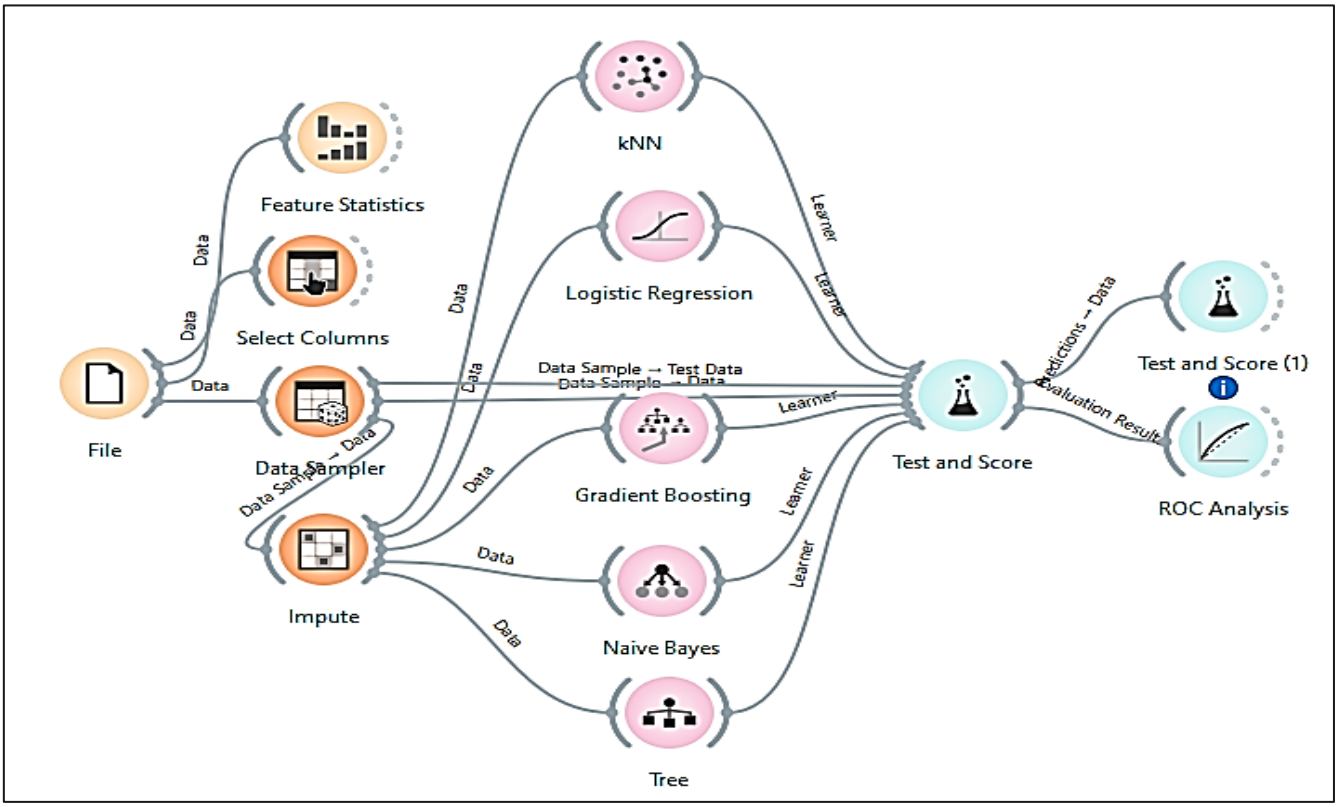

Figure 9: Orange Dashboard Machine Learning

Table III: Performance Comparison of Machine Learning Models(Model Used for Results)

S. NO.	Model	AUC	Accuracy (CA)	F1-Score	Precision	Recall	MCC
1.	Decision Tree	0.983	0.931	0.928	0.929	0.931	0.790
2.	Naïve Bayes	0.507	0.780	0.684	0.609	0.780	0.000
3.	Gradient Boosting	0.535	0.781	0.684	0.829	0.781	0.009
4.	k-Nearest Neighbors	0.788	0.798	0.757	0.771	0.798	0.275
5.	Random Forest	0.988	0.912	0.903	0.919	0.912	0.731

The key findings in Figure 9 show that the DT model outperformed all the others without a doubt, boasting the greatest accuracy of 0.931 and a well-balanced precision of 0.929 and recall of 0.931. This indicates that it can accurately forecast the outcomes of this categorization job far better than kNN or Random Forest. The Random Forest algorithm also performed well, achieving a high accuracy of 0.912 and fair metrics, but its results were slightly worse than those of the Decision Tree. With a low precision of 0.609, Naive Bayes, on the other hand, showed the least predictive potential, suggesting that it is not an appropriate algorithm. For this dataset, the Decision Tree is the most dependable classifier for forecasting survival outcomes since it has great accuracy and a well-balanced precision and recall performance, as shown by the results.

A. Confusion Matrix for K Nearest Neighbor

The confusion matrix of the k-Nearest Neighbor (kNN) model in Figure 10 illustrates the level of the model in the binary survival categorization test in terms of survival versus non-survival. According to the definitions of True Negatives and True Positives in the matrix, the model was correct in classifying 150,466 and 9,229 cases as not surviving and surviving, respectively. Nevertheless, the model did not succeed in its classifications. It has forecasted 34,673 real not survived cases as survived (False Positives) and 5,631 real survived cases as not survived (False Negatives). The overall amount of correct predictions (159,695) and false guesses (40,304) gives a definite upper hand to a correct classification. At a two way classification, the overall accuracy is approximately 79.8 percent. The survival rate is well predicted by this model with many successful predictions compared to failure predictions.

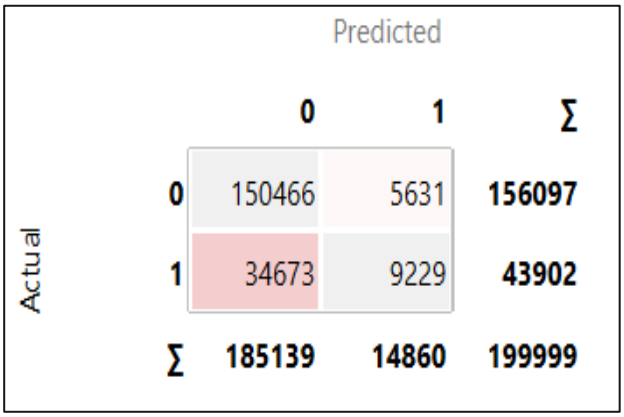

Figure 10: KNN Classifier Confusion Matrix of Lung Cancer Survival Prediction Indicating the Proportion of the Survived and Not Survived Classes that are True Positives, True Negatives, False Positives, and False Negatives

B. Confusion Matrix for Random Forest

According to Figure 11, the accuracy of the classification algorithm using the Random Forest is the percentage of cases that the classification algorithm correctly classified out of 155,716 cases, according to which 26,603 were correctly assigned to the having survived category (True Positives), and 26,603 were correctly assigned to the having not survived category (True Negatives). Nevertheless, misidentifications still influence the effectiveness of the model, with 381 successful survived cases of the model being predicted to not survive (False Negatives). Moreover, 17,299 real non survival cases were forecasted to be the survival (False Positives). The number of correct predictions was 182,319, and this is much more than the number of incorrect predictions (17,680). The Random Forest model in this case is strong and significant in terms of predictive power on survival outcomes, as indicated by the confusion matrix where the overall prediction of a binary categorization scenario is approximately 91.2%.

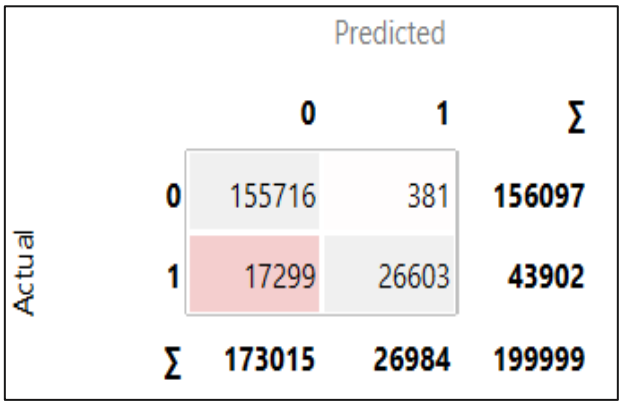

Figure 11: Random Forest Classifier Confusion Matrix of Lung Cancer Survival Prediction Indicating the Proportion of the Survived and Not Survived Classes that are True Positives, True Negatives, False Positives, and False Negatives

C. Confusion Matrix for Decision Tree

Figure 12 depicts the confusion matrix of the Decision Tree classifier that the researcher employed to determine the performance of the model in the prediction of survival in lung cancer.

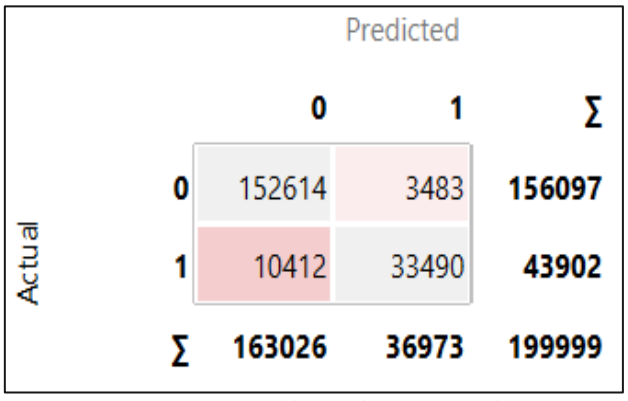

Figure 12: Decision Tree Classifier Confusion Matrix of Lung Cancer Survival Prediction Indicating the Proportion of the Survived and Not Survived Classes that are True Positives, True Negatives, False Positives, and False Negatives

The model appropriately categorized 152,614 non-survivors (true negatives) and 33,490 survivors (true positives). There were relatively few misclassifications, 3,483 non-survivors were wrongly predicted as survivors (false positives) and 10,412 survivors were wrongly predicted as non-survivors (false negatives). On the whole, the matrix indicates that the Decision Tree model would have an excellent balance of both correct survivor and non-survivor prediction with the proportion of errors being relatively low.

V. MODEL TRAINING AND EVALUATION CRITERIA

Each algorithm was trained using the preprocessed dataset, and performance was evaluated based on such measures as specificity, precision, recall, and F1-score. The classification exercise was primarily aimed at predicting two possible results.

1. Survived
2. Not Survived

To reduce bias and guarantee the reliability of the results, cross-validation techniques were used. The models' comparison analysis revealed which algorithm was most effective at predicting survival. This section describes the metrics also known as key performance indicators, or KPIs that will be used to assess the algorithm's output.

A. Accuracy

Number reflecting how well the predicted model performed. The accuracy formula shown in Eq. (1):

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

TP refers to where the model produces a positive class and the model correctly classifies the result as a positive. We call FP a False Positive outcome in a scenario whereby the model erroneously marks the positive class as a negative one. The outcome that the model predicted the negative class to be is what is referred to as the true negative, or TN. In the context of detection, a false negative, or FN is the name associated with negative response in a situation where a model thinks that the other category is wrong.

B. Precision

Precision is the proportion of cases that are accurately classified as positive. Specifically, if a model predicts positive numbers then the formula is shown in Eq. (2):

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

C. Recall

Recall is defined as the proportion of successfully recognized positives to all positives. This formula is the same as the sensitivity formula as shown by Eq. (3):

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

D. F1 Score

One metric for assessing a classification model's performance is the F1 score. It is a single metric that balances precision and recall by taking the harmonic mean of the two. It is shown by the Eq. (4).

$$F1\ Score = 2x\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (4)$$

The question of which algorithm is better may arise when precision and recall alone are not enough to assess performance, for as when one mining technique has higher accuracy but lower recall than another. This issue can be resolved by using the F1 score metric, which provides the mean of recall and precision. The F1 score is an industry standard for assessing a classification model's performance. The computation is displayed in Eq. (4). By combining recall and precision into a single score, it offers a fair assessment of a model's accuracy.

E. Matthews Correlation Coefficient (MCC)

Matthews Correlation Coefficient MCC is an equal-tailed estimate statistic that takes into account the four responses of a binary classification problem: true positives (TP) and true negatives (TN), false positives (FP) and false negatives (FN). In contrast to accuracy which can be misleading when applied on unbalanced datasets, MCC offers a more accurate estimation of the overall quality of the classifier. It is a value of -1 to +1. Overall mathematically shown in Eq. (5).

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (5)$$

The use of MCC is particularly important in this study because it provides a more reliable summary of model performance under class imbalance compared with accuracy alone.

F. Class Distribution Challenges and Mitigation Strategies

In this study, the dataset is imbalanced, with approximately 80% of patients in the Not Survived class and 20% in the Survived class. Such skewed class proportions can bias machine learning models toward the majority class. Such prejudiced allocations can bias the machine learning models so that they represent the majority, and consequently, they achieve a falsely high accuracy but not the outcomes as in the minority class. Such skew is particularly significant in medical prediction, where the error of false negativity (indicating that the patient survives when he or she does not) can be life-threatening to the clinical process. To address this weakness, in future studies, we ought to take into account the application of resampling techniques, such as Synthetic Minority Oversampling Technique (SMOTE), stratified cross-validation, or cost-sensitive learning, such that both classes are better represented. In addition, alternative metrics that are not pegged on the accuracy including MCC, F1-score and Precision-Recall AUC are to be targeted since they are more balanced and dependable assessments of the performance of classifiers on unbalanced medical data.

Although class imbalance is still a problem, skew was largely compensated for by automatically modifying class weights during training using ORANGE's built-in "Balance Data" widget. However, ORANGE does not support more sophisticated imbalance-handling techniques like SMOTE, cost-sensitive learning, or ensemble-based resampling. In order to further enhance minority-class recall and model fairness, these methods will be implemented in future versions of this work using a Python-based pipeline.

VI. INTERPRETATION OF TARGET FEATURE

The models were built to classify the patients based on whether they survived or not based on their demographics and clinical features. The target variable is survivorship, which can be of two categories: 0 (survived) and 1 (not survived). The image highlights 156,097 examples in an unbalanced sample, the bigger percentage of which were patients who did not manage to survive the outcomes of the lung cancer. The total number of the group that did not survive is 43,902, and there was the number of the group that did survive. Both groups comprise both male and female patients, as observed in Figure 13; however, the size of the classes is very different: the classes consist of the not survived group with almost 80% and the survived group with 20%. Such categorization division is justified by the chi-squared test ($\chi^2 = 2.94$, $p = 0.086$) that shows that survival status and gender are not statistically significantly associated. The rankings of the statistical features indicate that gender is not a major predictor of survival, whereas other aspects of the treatment, the level of cholesterol, smoking behavior, and type have more discriminatory features. Decision trees and random forests, among other machine learning models, yielded high AUC and accuracy in categorizing survival, implying that survival is important and the objective variable to be discriminated in the lung cancer data collection. Even though gender does not provide useful information to inform the discriminative variable in such an environment, the integration of clinical predictors and outcome characteristics allows the development of a successful predictive model to determine survival, shown in Figure 14. The boxplot shows how the gender groups are distributed in terms of survival. The graphical comparison reveals that there is no great asymmetry in the number of survivors and non-survivors of both sexes, and the trend is alike in both genders.

The clinical implications of prediction errors are that the level of risk associated with the various predictive errors is different. The most dangerous are false negatives, or the situation in the model where the patient can be predicted to survive but, in fact, the patient dies. Such errors can result in the clinicians giving less intensive monitoring, taking too long to escalate treatment, or even missing early palliative interventions. False positives, conversely - predicting a patient will die when the patient does not in fact die, might cause excessive resource allocation, but they are usually not as damaging as the missed risk-patients. The confusion matrices of each model thus serve as a good understanding of how the model would perform in the real clinical triage or risk-stratification environment. Practically, healthcare teams can change the decision thresholds to be more sensitive to the non-survived group so that high-risk patients can be detected even at the cost of a higher number of false positives. This is a threshold change that is typical of the clinical setting where reducing the number of missed critical cases is the priority.

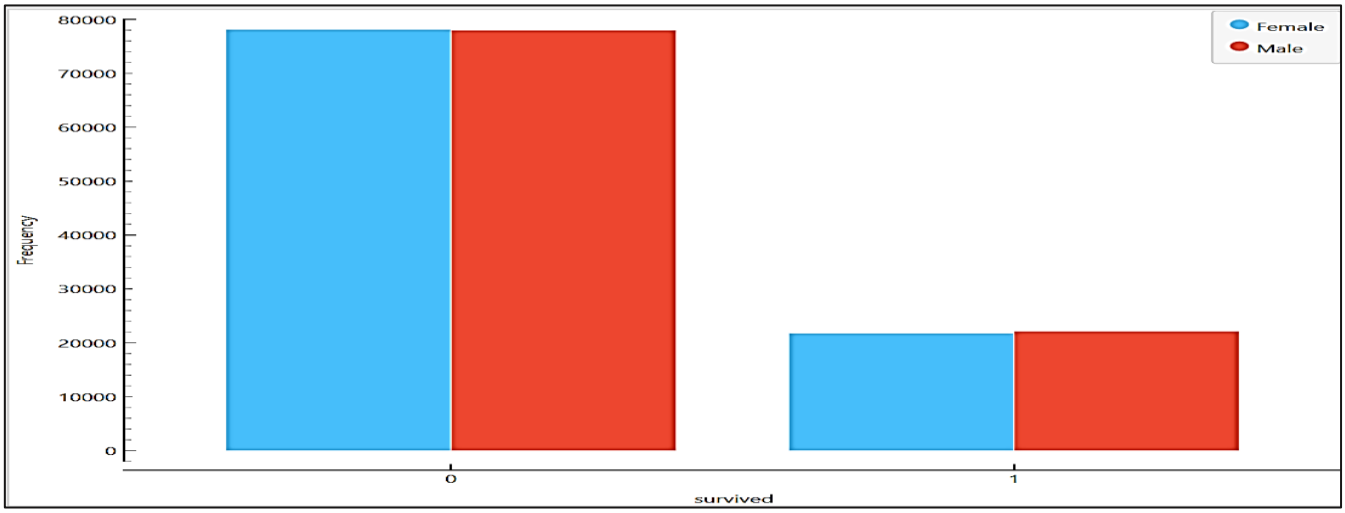

Figure 13: Survived or Not Survived Result

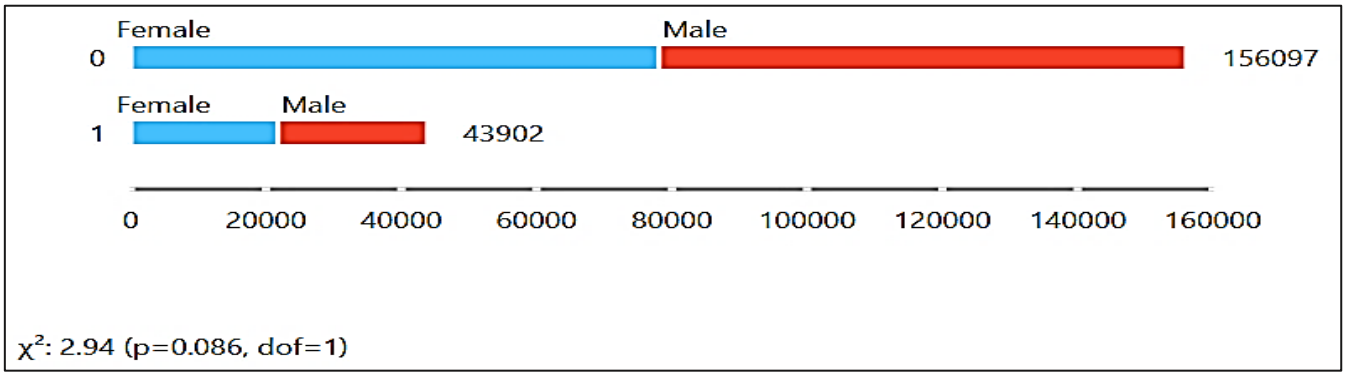

Figure 14: Boxplot Result Analysis on Outcome

VII. NOVELTY OF WORK

The originality of this study is that it targets survival as a predictive feature and simultaneously, integrates clinical outcome predictive modeling and strategic health care decision making based on the strength of Artificial Intelligence (AI) and Machine Learning (ML). Unlike the previous studies, which typically concentrate on the clinical, or algorithmic performance, this study cuts across the medical and management domains, and this is advantageous to both the patients and the health institutions leading to the provision of a comprehensive paradigm. It is based on a large scale Kaggle dataset of around 200,000 records of patients and covers a broad spectrum of demographic, lifestyle, and clinical characteristics, such as smoking status, BMI, comorbidities, treatment type, and gender. Further, despite the fact that the analysis of the various ML algorithms, including KNN, Decision Tree, Random Forest, Naive Bayes, and Gradient Boosting, is performed using the ORANGE data mining system, feature ranking using the assistance of Chi-square and ReliefF assists in exposing the most noteworthy factors that influence the survival outcomes. This two-pronged clinical administrative strategy makes sure that the research not only replaces an instrument of diagnosis with AI, but also makes it a strategic facilitator of healthcare systems. Such predictions can inform policy makers, hospitals and health insurance companies to reduce costs of treatment, distribute resources well and provide much equal and gender neutral healthcare.

In addition to technical performance, this research paper makes a direct linkage of the outputs of survival prediction to strategic decisions in healthcare. The proposed models will be able to inform medium-term planning in oncology departments regarding staffing, usage of chemotherapy chairs, scheduling of radiotherapy, and the allocation of ICU or high-dependency beds by identifying high-risk and lower-risk groups of survival.

VIII. MEDIUM-TERM IMPLICATIONS FOR HEALTHCARE MANAGEMENT

The most effective models can provide the survival probability as risk strata (scale: high, medium, and low probability of survival). Such strata may be actionable hospital planning and hospital management inputs. As an illustration, patients that were characterized as high-risk non-survivors could be given priority to more follow-up visits, multidisciplinary tumor boards, and early palliative care consultations. Aggregate forecasts within a group of patients at the operational level can be useful in supporting oncology units to forecast the demand of chemotherapy rooms, radiotherapy rooms and critical care rooms within a specific planning horizon. Moreover, such risk profiles can be utilized by insurers and policy makers to develop targeted intervention programs to patients with certain comorbidities and clinical weaknesses. This shows the usefulness of AI-based survival prediction as a clinical diagnosis tool but also as a health management and decision-making tool.

IX. CONCLUSION

This paper has shown that Artificial Intelligence (AI) and Machine Learning (ML) solutions could be successfully used to predict the outcomes of survival in lung cancer by considering patient demographics, medical history, and clinical characteristics. Several algorithms, such as Random Forest, Gradient Boosting, Decision tree, K-Nearest Neighbors (KNN) and Naive Bayes were made and evaluated on the basis of the ORANGE data mining platform. The Decision Tree was the most predictive to use among these. The ease with which Decision Tree models can be interpreted, and so can the Logistic Regression in the larger research, makes them exceptionally useful in clinical practice where transparency is of paramount importance. The results underline how AI-driven predictive models can be used to improve clinical decision making by assisting with early detection, risk assessment, and survival prediction in patients with lung cancer. In addition to clinical tasks, the incorporation of such predictive models into user-friendly software, such as ORANGE, has administrative value in the healthcare industry, such as enhancement of resource distribution, decrease in costs, and enhanced patient management approaches. Moreover, the paper underlines the promising potential of using a combination of clinical data, feature ranking algorithms (i.e. Chi-square and ReliefF), and machine learning algorithms to enhance survival prediction and facilitate evidence-based healthcare decision making. Further studies are needed to improve the accuracy and generalizability of the prediction by using larger and more heterogeneous datasets, genetic and imaging data, and overcome the issue of class imbalance by resampling and cost-sensitive learning. Besides that, the investigation of more complex deep learning designs including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks might also lead to the further improvement of predictive capability and clinical usability.

Moreover, we intend to work with nearby hospitals to validate the suggested models using actual registry and Electronic Health Record (EHR) data. In order to evaluate the survival forecasts' practical utility, calibration, and robustness in real-world healthcare settings, such clinical validation will be crucial.

X. LIMITATIONS

There are a number of limitations to this study. First, rather than coming directly from hospital Electronic Health Record (EHR) systems, the lung cancer dataset used in this work was obtained via a public Kaggle repository. As a result, there is no guarantee that the models can be applied to actual hospital populations, and the authenticity and clinical realism of the data may differ from actual patient registries. Second, imaging, genomic, and longitudinal follow-up data that could improve survival prediction are not included in the dataset. Third, the analyses have not yet been externally confirmed on other clinical cohorts and are based on a single dataset. These constraints suggest that rather than being instantly deployable models, the provided findings should be taken as an initial benchmark.

XI. ETHICAL CONSIDERATIONS IN AI-DRIVEN HEALTHCARE

Survival prediction with the help of AI and ML in lung cancer is an issue with critical ethical considerations. The information about patients should be managed according to the data protection laws and policies developed by institutional review boards. In the case of anonymized or public datasets, models that are trained on this type of data may incorporate demographic, access to care, or comorbidity biases. Regular auditing of model performance in subgroups and explainable approaches to enable clinicians to gain an insight into the decision to make a particular prediction are thus necessary. The use of the suggested models into practice in any real healthcare environment should be supported with proper governance, human regulation, and open communication with patients and other stakeholders.

Acknowledgment

The authors would like to express their sincere gratitude to the Department of Computer Science, Iqra University, Karachi, Pakistan, and the Department of Business Administration, Iqra University, Karachi, Pakistan, for providing academic guidance, institutional support, and a conducive research environment throughout the course of this study. The authors are also thankful to the Department of Computer Science, Karachi Institute of Economics and Technology (KIET), Karachi, Pakistan, for their valuable cooperation, technical support, and resources that contributed significantly to the completion of this research.

Authors Contributions

The authors equally contributed.

Conflict of Interest

The authors declare no conflict of interest and confirm that this work is original and not plagiarized from any other source.

Data Availability Statement

The testing data is available in this paper.

Funding

This research received no external funding.

References

[1] Raoof, S. S., Jabbar, M. A., & Fathima, S. A. (2020, March). Lung Cancer Prediction Using Machine Learning: A Comprehensive Approach. In *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)* (pp. 108–115). IEEE.

[2] Kadir, T., & Gleeson, F. (2018). Lung Cancer Prediction Using Machine Learning and Advanced Imaging Techniques. *Translational Lung Cancer Research, 7*(3), 304.

[3] Abdullah, D. M., Abdulazeez, A. M., & Sallow, A. B. (2021). Lung Cancer Prediction and Classification Based on Correlation Selection Method Using Machine Learning Techniques. *Qubahan Academic Journal, 1*(2), 141–149.

[4] Radhika, P. R., Nair, R. A., & Veena, G. (2019, February). A Comparative Study of Lung Cancer Detection Using Machine Learning Algorithms. In 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1–4). IEEE.

[5] Chaturvedi, P., Jhamb, A., Vanani, M., & Nemade, V. (2021, March). Prediction and Classification of Lung Cancer Using Machine Learning Techniques. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1099, No. 1, p. 012059). IOP Publishing.

[6] Patra, R. (2020, March). Prediction of Lung Cancer Using Machine Learning Classifier. In *International Conference on Computing Science, Communication and Security* (pp. 132–142). Springer.

[7] Pati, J. (2018). Gene Expression Analysis for Early Lung Cancer Prediction Using Machine Learning Techniques: An Eco-Genomics Approach. *IEEE Access*, *7*, 4232–4238.

[8] Dritsas, E., & Trigka, M. (2022). Lung Cancer Risk Prediction With Machine Learning Models. *Big Data and Cognitive Computing*, *6*(4), 139.

[9] Tuncal, K., Sekeroglu, B., & Ozkan, C. (2020). Lung Cancer Incidence Prediction Using Machine Learning Algorithms. *Journal of Advances in Information Technology*, *11*(2).

[10] Subramanian, R. R., Mourya, R. N., Reddy, V. P. T., Reddy, B. N., & Amara, S. (2020). Lung Cancer Prediction Using Deep Learning Framework. *International Journal of Control and Automation*, *13*(3), 154–160.

[11] Banerjee, N., & Das, S. (2020, March). Lung Cancer Prediction From a Machine Learning Perspective. In *2020 International Conference on Computer Science, Engineering and Applications* (ICCSEA) (pp. 1–5). IEEE.

[12] Sruthi, G., Ram, C. L., Sai, M. K., Singh, B. P., Majhotra, N., & Sharma, N. (2022, February). Cancer Prediction Using Machine Learning. In *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)* (Vol. 2, pp. 217–221). IEEE.

[13] Anil Kumar, C., Harish, S., Ravi, P., Svn, M., Kumar, B. P., Mohanavel, V., & Asfaw, A. K. (2022). [Retracted] Lung Cancer Prediction From Text Datasets Using Machine Learning. *BioMed Research International*, *2022*(1), 6254177.

[14] Lynch, C. M., Abdollahi, B., Fuqua, J. D., De Carlo, A. R., Bartholomai, J. A., Balgemann, R. N., et al. (2017). Prediction of Lung Cancer Patient Survival Via Supervised Machine Learning Classification Techniques. *International Journal of Medical Informatics*, *108*, 1–8.

[15] Mamatha, B., Rashmi, D., Tiwari, K. S., Sikrant, P. A., Jovith, A. A., & Reddy, P. C. S. (2023, August). Lung Cancer Prediction From CT Images Using Deep Learning Techniques. In *2023 Second International Conference on Trends in Electrical, Electronics, and Computer Engineering (TEECCON)* (pp. 263–267). IEEE.

[16] Farman, H., Khan, A. W., Ahmed, S., Khan, D., Imran, M., & Bajaj, P. (2024). Analysis of Supervised Machine Learning Techniques for Churn Forecasting and Component Identification in the Telecom Sector. *Journal of Computing & Biomedical Informatics*, *7*(01), 264–280.

[17] Farman, H., Khan, D., Hassan, S., Hussain, M., & Usmani, S. A. A. (2024). Analyzing Machine Learning Models for Forecasting Precipitation in Australia. *Journal of Computing & Biomedical Informatics*, *7*(01), 439–458.

[18] Farman, H., Hussain, A., Farman, A., & Irshad, S. (2025). Data-Driven Rainfall Prediction for Smart Agriculture: A Machine Learning Perspective. *Pakistan Journal of Engineering and Technology*, *8*(3), 31–41.

[19] Giner, J., Fernandez, P. R., Martinez, L. V., Bravo, A. R., Cano, O. C., Cobo, P. A., et al. (2025). Application of Artificial Intelligence: Machine Learning for Survival Prediction in Non-Small Cell Lung Cancer With Brain Metastases. *Journal of Thoracic Oncology*, *20*(3), S241.

[20] Vishwanathan Nair, V., & Soberanis, V. M. (2025). Lung Cancer Survival Prediction Using Machine Learning and Statistical Methods. arXiv e-prints *(arXiv:2510).*

[21] Farshchiha, S., Asoudeh, S., Kuhshuri, M. S., Eisaei, M., Azadi, M., & Hesaraki, S. (2025). A Comprehensive Analysis of Machine Learning-Based Methods for Lung Cancer Level Classification. *Intelligence-Based Medicine*, 100309.